Available online at www.sciencedirect.com
ScienceDirect
Genomics 89 (2007) 392–401
GENOMICS
www.elsevier.com/locate/ygeno

The frataxin-encoding operon of *Caenorhabditis elegans* shows complex structure and regulation

Rafael P. Vázquez-Manrique [a,b], Pilar González-Cabo [a], Inmaculada Ortiz-Martín [a], Sheila Ros [a], Howard A. Baylis [b], Francesc Palau [a,*]

[a] *Laboratory of Genetics and Molecular Medicine, Department of Genomics and Proteomics, Instituto de Biomedicina, CSIC, C/ Jaume Roig, 11, 46010 Valencia, Spain*
[b] *Department of Zoology, University of Cambridge, Downing Street, CB2 3EJ Cambridge, UK*

Received 15 June 2006; accepted 22 October 2006
Available online 15 December 2006

Abstract

The genome of the nematode *Caenorhabditis elegans* is unusual among eukaryotes, in that it contains operons. Approximately 15% of genes in the worm are clustered into groups of between two and eight genes, which are under the control of shared regulatory sequences. Polycistronic transcripts from such operons are trans-spliced, during transcription, to produce mature monocistronic messengers. The *C. elegans* frataxin gene, *frh-1*, is encoded in the operon CEOP2232. This is one of the largest operons identified thus far in the *C. elegans* genome. Here we describe in detail the structure of all of the coding units within this operon. The operon is composed of eight genes of a diverse nature, organized in a complex structure. We have produced transgenic strains carrying fusions between *gfp* and a number of genes from the operon. These constructs show complex differential expression patterns that suggest the presence of internal promoters and regulatory sequences in the operon. This organization would permit both coordinated expression and differential expression of the components of the CEOP2232 operon. The heterogeneity of the genes, and their complex expression patterns, suggests that the clustering of CEOP2232 is not due to a need for synchronized expression of genes involved in the same physiological pathway.

Keywords: Friedreich's ataxia; Frataxin; *frh-1*; *C. elegans*; Operon; Trans-splicing; Spliced leader; Promoter

Introduction

Some genes of the nematode *Caenorhabditis elegans* possess the extraordinary characteristic of being organized as polycistronic units or eukaryotic operons [1–3]. As in bacterial operons, these sets of genes share a common promoter and regulatory sequences that allow coordinated expression of the encoded proteins. In contrast with prokaryotic operons, *C. elegans* polycistronic transcripts are processed as they arise, producing mature monocistronic messengers that can then be transported to the cytoplasm and translated. This process involves another unusual feature of *C. elegans* gene expression, trans-splicing [4]. Approximately 70% of *C. elegans* genes are trans-spliced [5]. Trans-splicing uses a mechanism similar to that of cis-splicing but utilizes two sequences expressed from different loci, the spliced leader RNA (SL RNA) and the mRNA itself [4,6]. The pre-SL RNA donates the 22-nucleotide spliced leader (SL), which has a trimethyl G cap at its 5′ end that stabilizes the messenger and facilitates translation [7,8]. Trans-spliced leaders are spliced onto mRNAs containing an outron, an intron-like sequence that has only a 3′ acceptor site. There are two major types of trans-splicing in *C. elegans*: (1) trans-splicing of the spliced leader SL1 to standard mRNAs that contain an outron [4]; (2) trans-splicing of the spliced leader SL2 to mRNAs derived from polycistronic transcripts in which the outron is preceded by another coding sequence with a polyadenylation signal [5]. Consequently, SL1-mediated trans-splicing processes both single coding units and genes located immediately downstream of a promoter in an operon, whereas SL2-mediated trans-splicing processes downstream genes in operons [9,10]. In operons, the polyadenylation site of the

* Corresponding author.
E-mail address: fpalau@ibv.csic.es (F. Palau).

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.10.007

upstream gene and the trans-splicing site of the downstream gene are usually separated by between 30 and 360 bp, although in some cases the distance may be up to 1 or 2 kb [11]. SL2-mediated trans-splicing and polyadenylation of the upstream messenger are coupled [10,12]. Mutations in the polyadenylation site of the upstream gene reduce the efficiency of trans-splicing of the downstream gene, demonstrating that successful polyadenylation of a gene facilitates efficient SL2 trans-splicing of the downstream gene. This suggests that the efficiency with which a gene in an operon is processed will influence the expression of a downstream gene.

C. elegans operons contains an average of 2.6 genes, with 8 genes in the largest operon described thus far [9]. The orthologues of several genes involved in human disease are contained in operons [9]. This is the case for the *C. elegans* frataxin gene, *frh-1*, which is located within an unusually large operon, CEOP2232, which was predicted to contain 7 genes [9]. *frh-1* is the *C. elegans* homologue of *FXN*, deficiency of which causes Friedreich's ataxia in humans [13,14]. Human frataxin is a mitochondrial protein. The function of frataxin remains poorly understood, although several hypotheses have been proposed [15]. *FXN* is highly expressed in the heart and nervous system and in other tissues with a high energy requirement. Consistent with this expression pattern, its deficiency causes cardiomyopathy and neurodegeneration in patients. We have established a *C. elegans* model to study the function of frataxin and to test possible therapeutic treatments for Friedreich's ataxia [16]. An important part of this work is to determine when and where frataxin is expressed in *C. elegans*, thus allowing us to compare its expression with human frataxin. As *frh-1* is located within a large operon, it is therefore important to investigate the structure and regulation of this set of genes, as their expression may be relevant to the expression and function of the frataxin gene.

In this work, we present the first detailed description of a large operon in *C. elegans*. We have cloned and analyzed the mRNAs of all the genes in the CEOP2232 operon. We have shown that it is composed of eight rather than seven genes. We investigated the expression of the genes using transgenic strains carrying fusions of the reporter gene *gfp* (green fluorescent protein) to several genomic regions of the operon. These experiments revealed a complex picture in which frataxin is expressed in different tissues than other genes of the operon. Based on the results of these experiments, we propose the existence of multiple internal promoters and regulatory sequences within the frataxin-encoding operon.

Results and discussion

Structure of the operon CEOP2232

The *C. elegans* frataxin gene, *frh-1* [16], is contained within the operon CEOP2232 [9]. To characterize the expression of the frataxin gene, it was first necessary to understand, in detail, the structure of the CEOP2232 operon. When we started this work, there was little EST data for the genes in the operon. We therefore wished to determine, for each gene, its exact trans-splicing and polyadenylation sites together with its mRNA structure.

To determine trans-splicing sites, we performed RT-PCR using the sequences of the spliced leaders SL1 and SL2 as forward primers together with gene-specific reverse primers (Supplementary Table S1). This enabled us to determine the exact trans-splicing site(s) for each gene (Supplementary Table S2). Initially, we tested the most upstream gene in the predicted operon, *tsp-18*. We found that *tsp-18* was trans-spliced to both SL1 and SL2 (Table 1). The presence of SL2-spliced message (generally associated with internal genes within operons) suggested that *tsp-18* was not the first gene in the operon. The two genes upstream of *tsp-18* are *lin-4* and *F59G1.4*. *lin-4*, a miRNA [17], is not trans-spliced and so it is unlikely to be part of the operon. The predicted structure of *F59G1.4* placed its 3′ end 5 kb away from *tsp-18* (Fig. 1); however, by 3′-RACE we showed that the actual 3′ end of *F59G1.4* is in fact just 415 bp from the 5′ end of *tsp-18* (Table 1). The new predicted exons show a high level of similarity to the corresponding sequence of the genome of *C. briggsae* (Fig. 1). *lin-4* is thus now located in an intron of *F59G1.4* (Fig. 1), a common feature of miRNA genes [18]. *F59G1.4* is expressed only with the SL1-spliced leader (Table 1), suggesting that this is the gene leader in the CEOP2232 operon. Thus, the operon contains at least eight genes, making it one of the two largest operons described, thus far, in *C. elegans*.

We next performed similar experiments on the remaining six genes of the operon and determined their exact trans-splicing sites (Supplementary Table S2). To determine the polyadenylation sites for each mRNA, we performed 3′-RACE and cloned and sequenced the resulting PCR products (Supplementary Table S3). The distances between the polyadenylation and trans-splicing sites in the CEOP2232 operon range from 105 to 511 bp (Table 1), i.e., within the normal range of distances in *C. elegans* operons [11]. Finally, we amplified the sequence between the two mRNA ends of each gene and cloned and sequenced the resulting products to predict the structure of

Table 1
Trans-splicing sites for the genes of the operon CEOP2232

Gene	RT-PCR [a]		EST [b]		Intergenic distance [c]
	SL1	SL2	SL1	SL2	
F59G1.4	+	−	0	0	415
tsp-18	+	+	0	0	261
F59G1.8	+	+	0	0	202
vps-35	+	+	3	0	105
ctg-3 (F59G1.1b.1)	−	+	1	13	511
ctg-3 (F59G1.1b.2)	+	+	39	1	
ctg-3 (F59G1.1b.3)	+	−	2	0	
frh-1	+	+	0	2	328
ptp-2	+	+	1	0	345
vrk-1 (F28B12.3.1)	−	−	1	2	>8 kb [d]
vrk-1 (F28B12.3.2)	+	+	0	1	

[a] Presence (+) or absence (−) of SL1 or SL2 trans-spliced measured by RT-PCR (this work).
[b] Number of ESTs carrying SL1 or SL2 sequences shown in WormBase (www.wormbase.org release WS157).
[c] Distance between the polyadenylation site of a gene and the trans-splicing site of the downstream gene.
[d] Distance to the next predicted downstream gene (www.wormbase.org).

each gene. The proposed full structure of the operon derived from our results is shown in Fig. 2.

Downstream mRNAs are trans-spliced to both SL1 and SL2

Our analysis of the structure of CEOP2232 revealed that, with the exception of the gene leader *F59G1.4*, all of the genes in the operon were trans-spliced to both SL1 and SL2 (Table 1). However, the proportions of each mRNA that were trans-spliced to each spliced leader varied. In the case of *tsp-18*, the band of the SL1 product was stronger than that of SL2, suggesting preferential trans-splicing with that sequence. In contrast, *F59G1.8* produced a more intense band for the SL2 product (data not shown) and *vps-35* appeared to be equally expressed with SL1 and SL2. Analysis of the next gene in the operon, *cgt-3*, revealed a complex structure. When using the SL1 primer, we obtained products of several different sizes. When we cloned and sequenced these products, it became apparent that *cgt-3* has several alternative 5′ ends (Supplementary Table S2; Fig. 3). This is in agreement with the different predicted messengers described in WormBase for this gene. The relative strength of the SL1 and SL2 bands depends on the particular splice variant and is discussed in more detail below. For the frataxin gene, *frh-1*, we obtained more product when using SL2, indicating that the expression of this gene is predominantly with this sequence. Finally, both *ptp-2* and *vrk-1* showed equal and strong amplification with SL1 and SL2.

The operon CEOP2232 contains internal promoters

Several authors have proposed that *C. elegans* operons may contain internal promoters [19,20]. We have shown previously that the frataxin gene can be expressed from transgenes that do not contain the entire operon [16]. To do this, we produced constructs containing genomic regions, of a range of sizes, containing *frh-1* fused (in-frame) to the *gfp* gene, which we named *frh-1::gfp*1, *frh-1::gfp*2, and *frh-1::gfp*3 (Fig. 2b). Their expression patterns are detailed in Table 2. In summary, the three constructs showed different but specific and reproducible expression patterns, strongly suggesting that the region around *frh-1* contains several regulatory sequences. Importantly, these regulatory sequences promote the expression of the frataxin gene independently of the expression of the upstream genes, suggesting that frataxin expression is driven by internal promoters within CEOP2232.

To test for the presence of other internal promoters in the operon, we performed similar experiments on some of the other genes. First, we produced a construct of the 5′ region of the *cgt-3* gene, *cgt-3::gfp*. This gene has three alternative 5′ ends separated by relatively long noncoding sequences, which contain regions with a high level of similarity to the corresponding sequence from *C. briggsae* (data not shown). This structure suggests that these intronic regions may contain regulatory elements. The genomic region in the *cgt-3::gfp* construct overlaps substantially with constructs *frh-1::gfp*1 and 3 (Fig. 2b and Supplementary Fig. S1). Interestingly, this construct produced an expression pattern very similar to that of *frh-1::gfp*1 and 3, showing expression in the gut, pharynx, spermatheca, and anal cells (Table 2 and Figs. 4a–4c), indicating that the sequences that drive *frh-1* and *cgt-3* expression in these tissues are probably located in the *cgt-3* region (Supplementary Fig. S1). Moreover, there are several binding sites for GATA-like transcription factors in the overlapping region between *cgt-3::gfp*, *frh-1::gfp*1, and 3 (data not shown) [21]. However, in contrast to *frh-1::gfp*1 and 2, *cgt-3::gfp* did not show expression in head neurons, suggesting that regulatory sequences required for *frh-1* expression in the nervous system are located in a region close to or within the frataxin gene (Supplementary Fig. S1).

We next produced *gfp* fusion constructs for the *ptp-2* and *vrk-1* genes (*ptp-2::gfp* and *vrk-1::gfp*) (Fig. 2b and Supplementary Fig. S1), both of which produced consistent expression patterns (Table 2 and Figs. 4d–4h). *ptp-2::gfp* produced a complex expression pattern, which is different from that seen with *frh-1::gfp*, again indicating that the regulatory sequences that drive *ptp-2* expression are located close to the gene. Interestingly, Gutch and co-workers were able to rescue the phenotype of the *ptp-2(op194)* mutant, using the cosmid F28F12 [22], which contains only a part of CEOP2232, including *ptp-2*

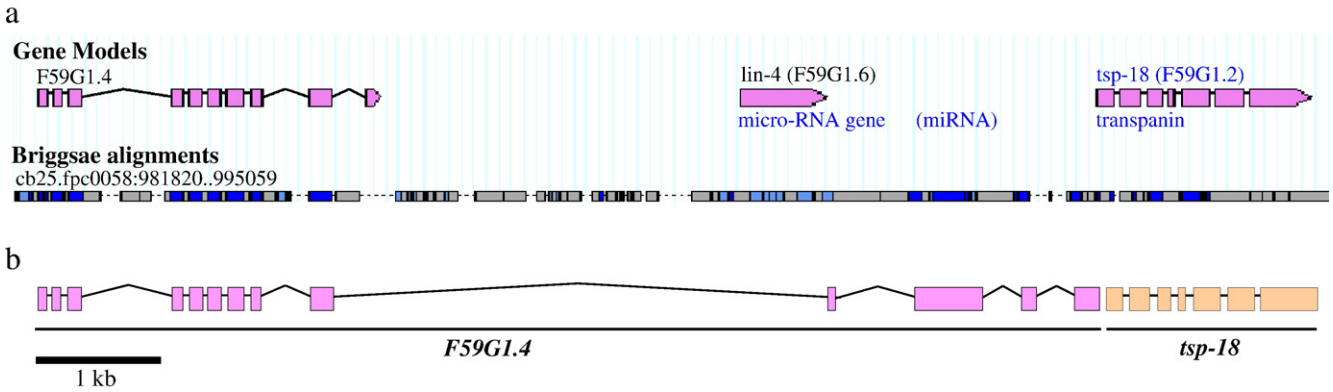

Fig. 1. Structure of the 5′ region of the CEOP2232 operon, showing genes *F59G1.4* and *tsp-18*. (a) Diagram of the region taken from WormBase showing the position and orientation of the predicted genes *F59G1.4*, *lin-4*, and *tsp-18*. Exons are represented in pink. Homology to the *C. briggsae* genome is shown at the bottom. Dark blue, light blue, and gray bars represent high, low, and no similarity, respectively. (b) Revised gene models and genomic arrangement, based on the analysis of cloned mRNA. Note that the *C. briggsae* sequence shows a high level of similarity in the region corresponding to the last three exons of the proposed *F59G1.4* gene.

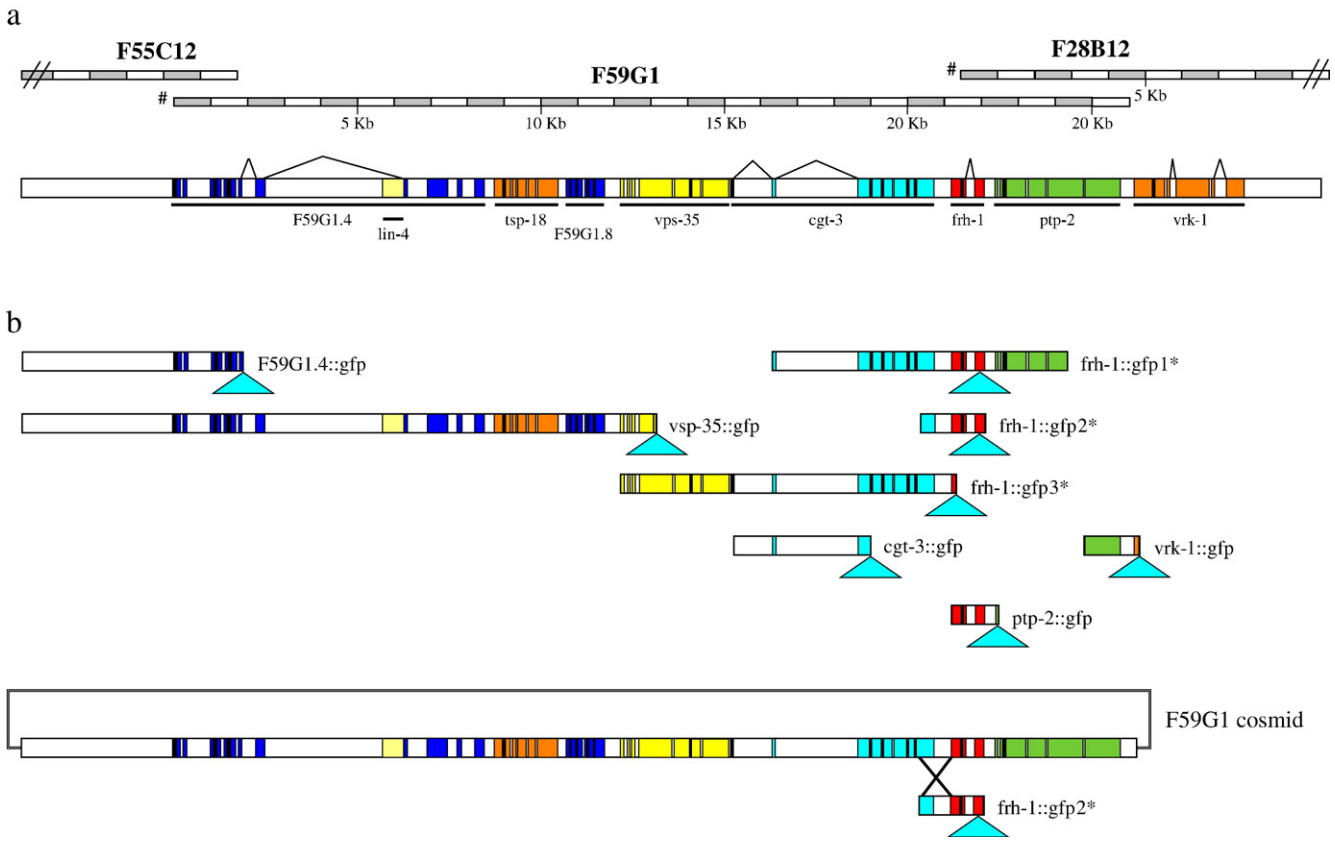

Fig. 2. The genomic structure of the CEOP2232 operon and *gfp* fusion constructs. (a) Colored boxes indicate sequences present in the mature mRNA (i.e., exons and UTRs). Each gene is distinguished by a different color. The sequenced region of the insert in cosmids that cover the region are shown at the top. (b) Diagram of constructs obtained either by fusion PCR (this work) or by cloning DNA fragments (marked with an asterisk) [15]. The putative homologous recombination event between *frh-1::gfp*2 and the cosmid F59G1 is shown at the bottom. The location of *gfp* is indicated by a blue triangle.

itself and a small region upstream. This means that sequences directly upstream of *ptp-2* are sufficient to drive its expression in the appropriate tissues, again indicating the existence of an internal promoter.

F59G1.4 and frh-1 are expressed differentially

The results presented above show that *frh-1* is expressed from internal promoters. However, we hypothesized that as the frataxin gene is part of an operon, we may also expect its expression to be co-regulated by sequences upstream of the 5′-most gene in the operon, *F59G1.4*. Hence, *frh-1* and *F59G1.4* should share similarities in their expression patterns. To investigate this question, we made a construct that included the whole region of the operon upstream of *frh-1* and in which *frh-1* was translationally fused to *gfp*. To achieve this, we took advantage of the ability to use naturally occurring homologous recombination to reconstruct genes in *C. elegans* [23–25]. We

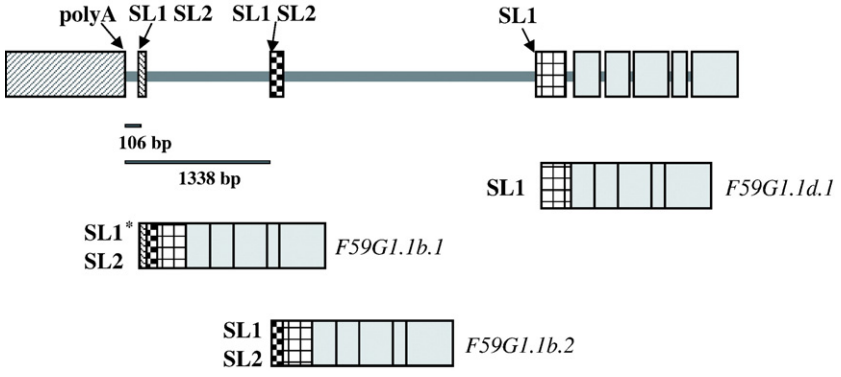

Fig. 3. Structure of the 5′ region of the *F59G1.1* gene (*cgt-3*) showing trans-splicing sites. Diagrammatic representations of the splicing variants together with the spliced leaders with which they are processed are shown at the bottom. Note that only the two exons closest to *vps-35* (106 and 1338 bp, respectively) are processed with SL2. The asterisk indicates that the presence of SL1 in this variant is rare (see Table 1).

Table 2
Expression patterns of *gfp* fusion constructs for genes in CEOP2232

Construct	Genome sequence [a]	5′ gene codons fused to *gfp*	Site of expression
F59G1.4::gfp	−3102–1642	1–209	Head neurons: inner labial neurons, amphids, phasmids
vps-35::gfp	−3102–12896	1–328	Head neurons: inner labial neurons, amphids, phasmids
cgt-3::gfp	15411–19054	1–55	Pharynx, spermatheca, gut, anal cells
*frh-1::gfp*1 [b]	15965–24000	1–135	Amphid and other head neurons, pharynx, gut, spermatheca
*frh-1::gfp*3 [b]	9090–21347	1–8	Pharynx, gut, spermatheca, muscle cells, anal cells
*frh-1::gfp*2 [b]	20654–22220	1–135	Amphid and other head neurons
*frh-1::gfp*2+F59G1 [b]	−3102–22220 [c]	1–135	Amphid and other head neurons, pharynx, gut, spermatheca, muscle cells, anal cells
ptp-2::gfp	22021–22637	1–8	Head neurons, spermatheca, gut, body wall muscles
vrk-1::gfp	24694–4459 [d]	1–8	Head neurons, spermatheca, gut, pharynx, vulva, phasmids

[a] Nucleotide number within the F59G1 cosmid (Accession No. U53332). Under Genome sequence, −3102 indicates the 5′ nucleotides upstream of the first gene in the operon, *F59G1.4*.
[b] Described in [15].
[c] Construct obtained by homologous recombination.
[d] Nucleotide number at 3′ corresponds to position in cosmid F28B12 (Accession No. U29537).

produced transgenic animals carrying extrachromosomal arrays derived from an equimolar mixture of the construct *frh-1::gfp*2 and the cosmid F59G1. F59G1 contains most of the operon and some upstream DNA [16] (Fig. 2b). *frh-1::gfp*2 expresses *gfp* only in a few head neurons (Table 2). In contrast, the transgenic lines carrying arrays made of *frh-1::gfp*2 and the F59G1 cosmid mimic the expression pattern produced by the constructs *frh-1::gfp*1 and *frh-1::gfp*3. This suggests that homologous recombination fused both molecules, thus assembling the operon from upstream of its start site down to *frh-1* fused to *gfp*. We can assume that this is a good representation of the real expression pattern of the frataxin gene. These strains showed expression of *gfp* in several tissues: muscles, gut, pharynx, spermatheca, and head neurons. To compare this expression pattern with that resulting from the putative promoter at the start site of the operon, we produced two *gfp* fusion constructs. The first contained the 5′ upstream region of the operon fused to *gfp*, in-frame with the eighth exon of *F59G1.4*, the operon gene leader (Fig. 2b and Supplementary Fig. S1). Animals carrying this construct show a different expression pattern from that of *frh-1*. The pattern was predominantly neuronal (Fig. 5a, and Table 2) and, in contrast with *frh-1* expression, was not observed in the gut, pharynx, or spermatheca. The nature of the neuronal expression was also different, as *frh-1* is expressed in the amphid cells and other neurons around the posterior part of the pharynx (Fig. 5a and Table 2), whereas *F59G1.4* was expressed in amphid and labial neurons (Fig. 5b and Table 2) and posterior phasmid neurons. The second construct contained the 5′ upstream region of the operon to the fourth exon of the second gene in the operon, *vps*-35, again fused to *gfp* (Fig. 2b and Supplementary Fig. S1). The expression of this construct matched the expression of the *F59G1.4* gene (Table 2), suggesting that the expression of both genes is under the control of the same regulatory elements. Hence, the genes situated at the 5′ region of the operon (at least *F59G1.4* and *vps-35*) are expressed in different cells than *frh-1*, suggesting that the promoter located upstream of *F59G1.4* does not play a substantial role in regulating the expression of some of the genes within the operon.

The presence of internal promoters in the operon CEOP2232 may contribute to the trans-splicing of SL1 to downstream genes

As mentioned previously, most of the transcripts in CEOP2232 are spliced to both SL1 and SL2. The presence of internal promoters in the operon would explain the presence of SL1-spliced internal transcripts. This hypothesis has been suggested previously [19]. However, it has also been proposed that the degree of coupling between adjacent genes within an operon may influence the relative levels of SL1- and SL2-spliced message, so that a reduction in coupling efficiency to SL2 may lead to SL1 splicing. In this regard, *cgt-3* may be informative. As described above, the first exon of *cgt-3* is located just 100 bp downstream of *vps-35*. The clones isolated in this work containing this exon were trans-spliced to SL2. Furthermore, of 14 ESTs currently reported that contain a spliced leader sequence, only one (yk1437e11) is trans-spliced to SL1 (Table 1). The *cgt-3* isoform produced by the expression of this exon is *F59G1.1b.1* (Fig. 3). Clones obtained for the variant that starts with the next exon downstream, *F59G1.1b.2*, are trans-spliced with both SL1 and SL2, although ESTs are very SL1 biased (Table 1). Finally, based on our data and ESTs, the variant *F59G1.1d.1* (beginning at the third exon) is trans-spliced exclusively to SL1. These results may be explained by different coupling efficiencies and by the presence of internal promoters. The most upstream spliced variant begins with an exon that is very close to *vsp-35*, which couples efficiently to *vps-35* processing and is thus predominantly spliced to SL2, with only rare "aberrant" splicing SL1. The most downstream variant is not coupled to *vps-35* processing and is transcribed from a promoter in the large upstream intron and is thus SL1 trans-spliced. The middle splice variant couples to *vps-35* polyadenylation, thus resulting in some SL2 trans-splicing, but its distance from *vps-35* (approximately 1.3 kb) makes this somewhat inefficient, leading to more SL1 splicing. Alternatively, another promoter may lie upstream of this variant and this may drive the SL1 trans-spliced mRNAs.

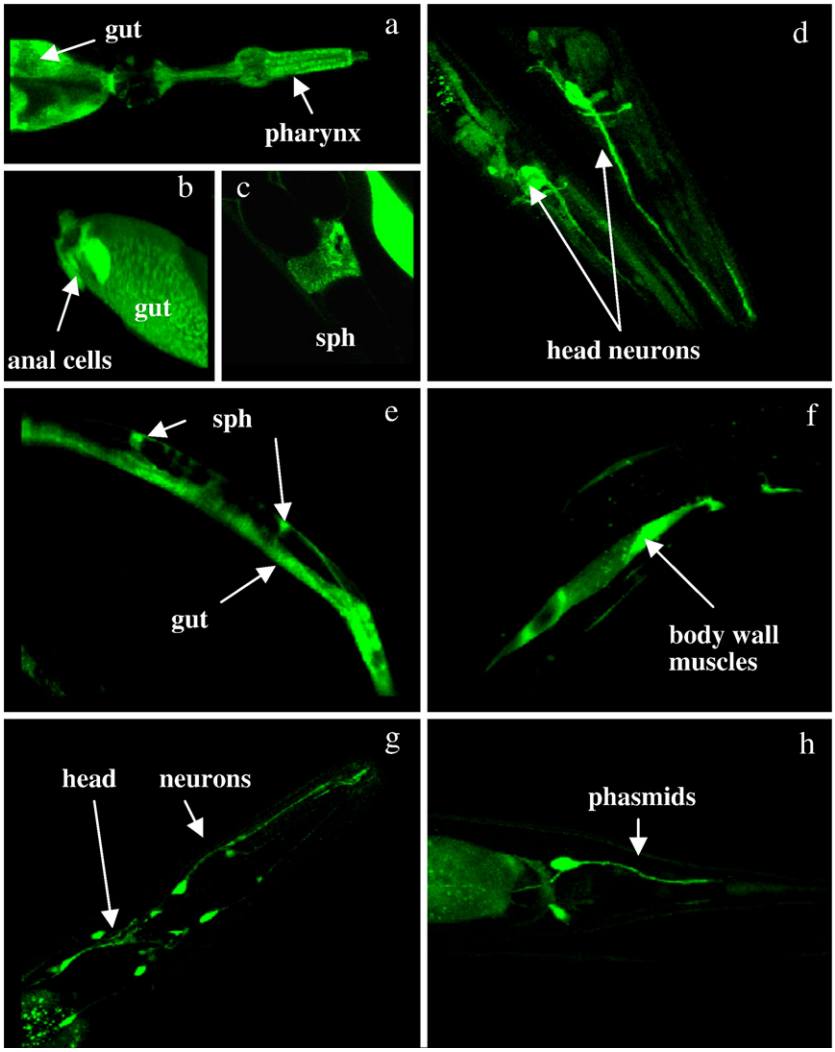

Fig. 4. GFP expression driven by autonomous regulatory sequences in transgenic worms carrying constructs *F59G1.1::gfp*, *ptp-2::gfp*, and *vrk-1::gfp*. (a–c) Expression of *F59G1.1::gfp* in pharynx and gut (a), anal cells (b), and spermatheca (sph) (c). (d–f) *ptp-2::gfp* produced fluorescence in some head neurons (d), gut and spermatheca (e), and body wall muscle (f). (g–h) *vrk-1::gfp* was expressed in several cells, which included head neurons (g) and phasmids (h).

To further investigate the balance of splicing of SL1 and SL2 due to internal promoters and coupling, we performed an experiment in which we could compare the endogenous trans-splicing of *frh-1* with the trans-splicing of this gene present in chromosomal arrays. To accomplish this, we used worms carrying the construct *frh-1::gfp* 2 (Figs. 2a and 6a). We carried

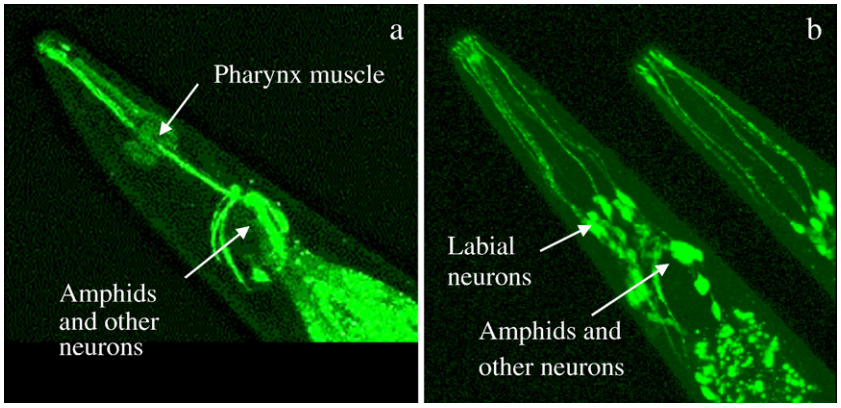

Fig. 5. Differential expression of *frh-1* and *F59G1.4*. (a) Representative confocal image of animals carrying the array composed of *frh-1::gfp*2 and the cosmid F59G1. (b) Image of animals carrying the construct *F59G1.4::gfp*.

out RT-PCR using SL1 or SL2 forward primers on RNA obtained from these worms, followed by Southern blot analysis of the products. To amplify endogenous transcripts from the genes, we used a gene-specific reverse primer (Supplementary Table S1), and to specifically amplify transcripts from the transgenes, we used a primer in the *gfp* sequence. The results show that endogenous expression of *frh-1* is mostly SL2 spliced, presumably because expression is coupled to polyadenylation of the upstream gene (Fig. 6b). In contrast, the expression of the extrachromosomal constructs, which is possible only through internal promoters, shows higher SL1 splicing (Fig. 6b). These data suggest that internal promoters can be a source of SL1-spliced mRNA for genes located downstream in an operon that would otherwise be trans-spliced to SL2.

The CEOP2232 operon comprises a heterogeneous set of genes, but its structure is highly conserved through the genus Caenorhabditis

It has been demonstrated that, in some cases, the genes located in *C. elegans* operons are implicated in the same physiological pathway or function, as is common in prokaryotic

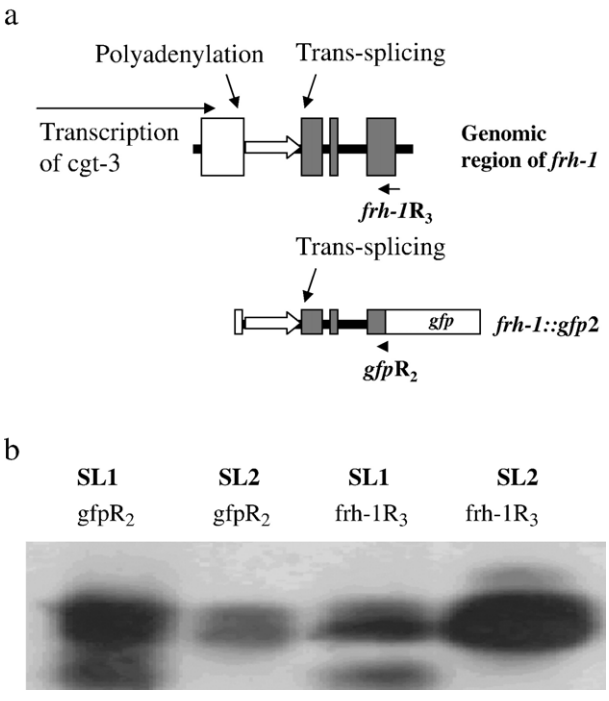

Fig. 6. Trans-splicing in endogenous and transgenic *frh-1*. (a) Genomic region of *frh-1* and the construct used to investigate the trans-splicing of this gene. The reverse primers used are shown under each structure. *gfp*R_2 is specific to *gfp* and thus will only give products from expression of the transgene. Only the endogenous expression of *frh-1* can be coupled to the processing of *cgt-3*. This will increase the contribution of SL2 trans-splicing. In contrast, expression from the transgene can only result from internal promoters (white arrow upstream of *frh-1*), thus producing SL1 trans-spliced mRNAs. (b) Southern blot showing the products of RT-PCR for transgenic expression (the leftmost two lanes) or endogenous expression (the rightmost two lanes) of *frh-1*, in the transgenic animals. Expression from the transgenic array produces more SL1-spliced *frh-1* product, whereas the autonomous expression of *frh-1* is associated mostly with SL2 [15].

operons [26–28]. Blumenthal and Gleason demonstrated the existence of several operons containing genes for proteins with clearly related functions and discussed the possibility that these genes are co-regulated to synchronize their expression [3]. Furthermore, this arrangement has made it possible to predict the function of a human gene, based on the presence of a *C. elegans* homologue in an operon containing a gene of known function [29]. Hence, operons can be used as a tool to identify partner proteins [30], something that would contribute to our understanding of frataxin function. Although the function of frataxin is not well understood, it is known to be mitochondrial and a range of mitochondria-related roles have been suggested for it, including the following: free radical detoxification, heme biosynthesis, and energy conversion [15]. Thus, if any of the proteins of CEOP2232 are related to frataxin in function, we might expect that these proteins would also be mitochondrial or would have functional properties related to proposed frataxin functions. We surveyed the nature of the genes in the operon. There are two genes, *tsp-18* and *F59G1.8*, with no clear homologues in other organisms. *F59G1.8* has no known domains, so its function cannot be predicted. *tsp-18* is a member of the tetraspanin superfamily, which regulates cell migration, fusion, and signaling events [31]. Neither of these proteins is predicted to be mitochondrial by the Mitoprot II program [32]. *F59G1.4* has a very low level of similarity to a human protein of unknown function, ARMC9 [33], but does have a predicted probability of 73.3% that it is mitochondrial (using Mitoprot II). The remaining proteins have clear homologues: *vps-35* is an orthologue of the yeast vacuolar protein sorting-associated protein 35; *cgt-3* is similar to ceramide glycosyltransferases involved in the synthesis of membrane glycolipids; *ptp-2* is a nonreceptor tyrosine phosphatase involved in *C. elegans* oogenesis [22]; and vrk-1 is a protein kinase that is essential for *C. elegans* viability (www.wormbase.org). Therefore, the CEOP2232 operon contains a heterogeneous set of genes, with no clear pattern of related function.

Given the apparent heterogeneity of the genes in the operon, we wondered to what extent this gene arrangement is conserved. We surveyed the structure of the operon in the genus *Caenorhabditis*, by comparing CEOP2232 to the *C. briggsae* and *C. remanei* genomes. Six of the eight genes (*F59G1.4*, *vps-35*, *cgt-3*, *frh-1*, *ptp-2*, and *vrk-1*) are highly conserved and are located in the same order and separated by similar distances in both of these species (data not shown). *tsp-18* is also present in a similar position, but is more diverged (Supplementary Fig. S2a). The *F59G1.8* sequence also shows some degree of similarity between *C. elegans*, *C. briggsae*, and *C. remanei* (Supplementary Fig. S2b) and is also located in a similar position. Hence, the structure of the operon is conserved among the three species. The location of the *C. elegans* frataxin gene within a large operon made up of such a heterogeneous set of genes raises questions about the function of this organization. One possibility is that, although these genes do not show any apparent link to frataxin function, they may nevertheless be directly or indirectly involved in the regulation of frataxin production or function. Alternatively, the genes may not be functionally related, but may instead share common regulatory

requirements and thus common regulatory elements. It has been proposed that small genomes, such as that of *C. elegans*, can favor compaction of genes and hence the emergence of operons, as genes evolve to share promoters and other regulatory sequences [3]. However, the complex nature of the regulatory elements in COP2232 make a simple model of such co-regulation problematic in this case.

CEOP2232 shows differential selective pressure along its structure: tsp-18 and F59G1.8

Despite the high degree of conservation of the structure of the CEOP2232 operon within the genus *Caenorhabditis* and of many of the genes in the operon, *tsp-18* and *F59G1.*8 are not highly conserved in sequence. In addition, they both appear to be expressed at low levels. cDNAs of *tsp-18* were difficult to detect and there are no ESTs for this gene (www.wormbase.org release WS157). We isolated one cDNA, which did not contain the first two predicted exons (www.wormbase.org). *F59G1.8* also appears to be present at low levels. We consistently obtained very low yields in PCR assays using several different sets of primers along the predicted gene and, again, there are no ESTs for *F59G1.8* (www.wormbase.org release WS157). We obtained a single cDNA clone of *F59G1.8* with an unpredicted exon between predicted exons 3 and 4. However, exon 4 was 23 bp longer than predicted, leading to an mRNA in which the reading frame in the next to last exon is disrupted. This clone may be the result of an aberrant splicing event, but another possible explanation is that *F59G1.8* is a pseudogene. In addition to apparently very low levels of expression, *F59G1.8* shares very little sequence similarity with the closely related species *C. briggsae* and *C. remanei* [34] (Supplementary Fig. S2), suggesting that this gene has not been subject to selective pressure. If *F59G1.8* is a nonfunctional gene, why is it present within the operon? It might be argued that it is required to maintain the functional architecture of the operon and thus correct expression of downstream genes.

Conclusions

In this work, we describe the structure and regulation of the expression of the frataxin-encoding operon, CEOP2232. We provide data supporting the presence of internal and autonomous promoters within the operon. This hypothesis has been suggested previously [19]. It has also been proposed that differential expression levels of genes expressed from the same promoter may be achieved by posttranscriptional regulation [11]. However, this cannot explain all of our results, as some of the constructs discussed in this work do not contain the coding sequence of the gene of interest, but only *gfp* fused to their first codons (*cgt-3::gfp* and *frh-1::gfp*3). The internal promoters in CEOP2232 may serve to produce differential expression of genes through transcriptional regulation. We show that internal promoters may be a source of SL1-mediated trans-splicing, as suggested previously [19]; however, it also seems likely that some of the SL1 splicing, especially that observed in predominantly SL2-spliced mRNAs, results from poor coupling rather than transcription from internal promoters. Our results show that there are at least two groups of co-regulated genes. The first block includes the genes in the 5′ region of the operon, from *F59G1.4* to *vps-35*, and possibly the *F59G1.1b.1* form of *cgt-3*. Their expression is driven by a promoter located upstream of the whole operon, resulting in a similar neuronal pattern of expression. The second group may contain the genes downstream of *vps*-35, expressed from a promoter located in the interexon regions at the 5′ end of *cgt-3*, which produces expression in tissues other than neurons, such as gut, spermatheca, and pharynx. This region regulates, at least, *cgt-3* and *frh-1*. It is possible that a third block of expression exists, as *ptp-2* and *vrk-1* seem to have autonomous promoters as well. Why is there such a striking difference in the expression of the *F59G1.4*-associated group from that of the frataxin-associated group? It is tempting to argue that these two groups of genes are the remains of two ancient operons that became combined in the same polycistronic unit through *cgt-3*, which links the expression of *vsp-35* to that of the downstream genes. Thus, the presence of internal promoters within the CEOP2232 operon may be a vestige of the autonomy of the downstream genes, now integrated in a polycistronic structure. Blumenthal and coworkers have discussed the possibility that polycistrons may be stable structures because downstream genes in an operon lose their autonomy, since their expression depends on the promoters of upstream genes [3,11]. In the case of CEOP2232, selective pressure may also preserve internal regulatory sequences, as the expression of its constituent genes is different. Analysis of these genes in more distantly related species may shed further light on these evolutionary questions.

Finally, is CEOP2232 a peculiarity in the *C. elegans* genome or are there more examples of such complex operons? The existence of internal promoters providing a finer regulation of some genes of the ABC transporter family has been suggested [20]. Another example of a potential internal promoter within an operon has been shown by Xue and co-workers [35], who described the expression of *mtm-6*, the last gene contained in the 5-gene operon CEOP3784. A construct containing 3.3 kb of genomic DNA directly upstream of the first exon of *mtm-6* drives *gfp* in the intestine [35] (Supplementary Fig. S3). This genomic fragment does not contain the whole 5′ region of the operon, so that this expression is not driven by a promoter upstream of the first gene of the polycistronic unit. Like *cgt-3*, *mtm-6* has a complex structure in its 5′ region that produces alternative splicing (Supplementary Fig. S3). We produced an additional *mtm-6* construct, using the genomic region containing the 5′ region of the gene, including a large intron (4 kb) that we suspected could contain another internal promoter. This construct showed neuronal expression (Supplementary Fig. S3), which is in contrast with the intestinal expression described previously for *mtm-6* [35]. Hence, CEOP3487 also contains internal promoters that regulate *mtm-6* in a complex manner. We propose that common regulatory sequences within operons such as CEOP2232 and CEOP3784 may provide coordinated regulation of constituent genes, but additional control through internal promoters provides further fine regulation, and that this characteristic may be found in other *C. elegans* operons.

Materials and methods

Strains and worm culture

Worms were cultured using standard techniques and media [36,37]. We used the wildtype strain Bristol (N2), supplied by the *Caenorhabditis* Genetics Centre (University of Minnesota, MN), for all experiments. All strains were maintained at 20°C.

Reverse transcription and PCR

Worms were grown on standard NGM plates. RNA was isolated from these worms using an RNeasy kit (Qiagen), following the manufacturer's method for animal tissues. To remove DNA, RNA was treated with DNase I (Roche) for 20 min at 37°C followed by inactivation at 75°C for 20 min. For cDNA synthesis, 5 μg of total RNA was reverse-transcribed using random hexamers as primers and 4 U of MMLV reverse transcriptase (Invitrogen). Subsequent PCRs were performed using HiFi DNA Polymerase (Roche), according to the manufacturer's instructions.

Southern blot analysis

We prepared total RNA from the strains carrying the construct *frh-1::gfp*2 and performed RT-PCR as described above, using the primers described under Results. The products of these reactions were separated on a 0.8% agarose gel and then transferred to a nylon membrane (Roche), using a standard procedure [38]. The filters were probed with the entire F59G1 cosmid labeled with digoxigenin using random primers, following the procedure recommended by the manufacturer (Roche). Hybridization was carried out overnight at 65°C and the membrane was washed at high stringency: 0.1× SSC, 0.1% SDS at 65°C (where 1× SSC is 15 M NaCl and 0.015 M sodium citrate). Detection was performed using an alkaline phosphatase anti-digoxigenin antibody and CSPD Star as the enzyme substrate (Roche) and visualized using X-ray film.

Gene cloning and sequencing

To identify the ends of each gene, we first tested for trans-splicing to SL1 and SL2 using a gene-specific oligonucleotide as the reverse primer (Supplementary Table S1). The 3′ ends were determined using 3′-RACE, carried out as described previously [39]. Products were cloned into pGEM-T (Promega) and then sequenced. Having established each end, we used this information to amplify full-length cDNAs. These products were cloned and sequenced, as described above, to obtain the whole structure of each gene. All transcripts were easily amplified, apart from those corresponding to *tsp-18* and *F59G1.8*, which were all present at low levels and were detected by Southern blot. The sequences obtained were deposited with the GenBank Data Library under the following accession numbers: *F59G1.4*, DQ178637; *tsp-18*(*F59G1.2*), DQ178635; *F59G1.8*, DQ178638; *vps-35*(*F59G1.3*), DQ178636; *cgt-3*, DQ178632-DQ178634 (*F59G1.1*) ; *frh-1*(*F59G1.7*), AY048153 [16]. The sequences obtained for *ptp-2*(*F59G1.5*) and *vrk-1*(*F28B12.3*) matched those of the *F59G1.5.1* and *F28F12.3.2* predicted models (www.wormbase.org, ws157).

Generation of gfp fusions

To investigate the expression pattern of genes, we produced PCR products in which *gfp* was fused to the genomic region to be investigated and in-frame with the gene of interest. The fusion of genomic and *gfp* sequences was achieved by PCR fusion [40]. Briefly, we generated two PCR products: in PCR 1, we amplified the gene of interest with appropriate upstream genomic sequences (see Supplementary Table S1 for the primers used), and in PCR 2, the *gfp* coding sequence plus the 3′-UTR from *unc-54* gene were amplified from pPD95.75 (http://www.ciwemb.edu/). The 3′ primer used in PCR 1 had a 24-nt overhang that matched *gfp* in pPD95.75. Fusion transgenes were then obtained by nested PCR using as template a mixture of the products from PCRs 1 and 2 at 20–50 ng/μl each (Supplementary Table S1). For this reaction, nested primers downstream of the 5′ primer in PCR 1, specific for each gene, and a specific primer for the *gfp* cassette were used.

Production of transgenic animals and microscopic observation

The transgenic strains were generated by microinjection [41]. We injected the *gfp* fusion constructs at a concentration of 20–50 ng/μl into wildtype worms. We used the dominant marker *rol-6(su1006)* (plasmid pRF4) [42] at a concentration of 100 ng/μl, to select transgenic animals. We examined at least two independent transgenic lines for every construct injected. Images were obtained from worms anesthetized in 1 mM levamisole using a Leica SP laser scanning confocal microscope (Leica Microsystems).

Acknowledgments

We thank N. Gower and D. Walker, for helpful comments and suggestions. We are grateful to the *Caenorhabditis* Genetic Centre for providing some of the strains used in this work and to A. Fire for supplying vectors used to produce transgenic animals. The collaboration between the Baylis and Palau groups was supported by a Biomedical Research Collaboration Grant from the Wellcome Trust (Grant 069898). This work was supported by the Spanish Ministry of Education and Science (Grant SAF2003-00135), the Instituto de Salud Carlos III (Grant G03/056 for the Spanish Network on Cerebellar Ataxias), and the Generalitat Valenciana (Grant GRUPOS03/015). H. A. Baylis is an MRC Senior Fellow.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2006.10.007.

References

[1] T. Blumenthal, Gene clusters and polycistronic transcription in eukaryotes, BioEssays 20 (1998) 480–487.
[2] T. Blumenthal, Operons in eukaryotes, Brief Funct. Genomic Proteomic 3 (2004) 199–211.
[3] T. Blumenthal, K.S. Gleason, *Caenorhabditis elegans* operons: Form and function, Nat. Rev. Genet. 4 (2003) 112–120.
[4] M. Krause, D. Hirsh, A trans-spliced leader sequence on actin mRNA in *C. elegans*, Cell 49 (1987) 753–761.
[5] K.E. Hastings, SL trans-splicing: Easy come or easy go? Trends Genet. 21 (2005) 240–247.
[6] X.Y. Huang, D. Hirsh, A second trans-spliced RNA leader sequence in the nematode *Caenorhabditis elegans*, Proc. Natl. Acad. Sci. USA 86 (1989) 8640–8644.
[7] S. Lall, C.C. Friedman, M. Jankowska-Anyszka, J. Stepinski, E. Darzynkiewicz, R.E. Davis, Contribution of trans-splicing, 5′-leader length, cap-poly(A) synergism, and initiation factors to nematode translation in an *Ascaris suum* embryo cell-free system, J. Biol. Chem. 279 (2004) 45573–45585.
[8] P.A. Maroney, J.A. Denker, E. Darzynkiewicz, R. Laneve, T.W. Nilsen, Most mRNAs in the nematode *Ascaris lumbricoides* are trans-spliced: A role for spliced leader addition in translational efficiency, RNA 1 (1995) 714–723.
[9] T. Blumenthal, D. Evans, C.D. Link, A. Guffanti, D. Lawson, J. Thierry-Mieg, D. Thierry-Mieg, W.L. Chiu, K. Duke, M. Kiraly, S.K. Kim, A global analysis of *Caenorhabditis elegans* operons, Nature 417 (2002) 851–854.
[10] J. Spieth, G. Brooke, S. Kuersten, K. Lea, T. Blumenthal, Operons in *C. elegans*: Polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions, Cell 73 (1993) 521–532.
[11] T. Blumenthal, Trans-splicing and operons, in: The *C. elegans* Research

Community (Eds.), WormBook: The Online Review of *C. elegans* Biology, 2005, doi/10.1895/wormbook.1.5.1, 2005, http://www.wormbook.org.

[12] S. Kuersten, K. Lea, M. MacMorris, J. Spieth, T. Blumenthal, Relationship between 3′ end formation and SL2-specific trans-splicing in polycistronic *Caenorhabditis elegans* pre-mRNA processing, RNA 3 (1997) 269–278.

[13] V. Campuzano, L. Montermini, M.D. Molto, L. Pianese, M. Cossee, F. Cavalcanti, E. Monros, F. Rodius, F. Duclos, A. Monticelli, et al., Friedreich's ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion, Science 271 (1996) 1423–1427.

[14] F. Palau, Friedreich's ataxia and frataxin: Molecular genetics, evolution and pathogenesis, Int. J. Mol. Med. 7 (2001) 581–589.

[15] M. Pandolfo, Frataxin deficiency and mitochondrial dysfunction, Mitochondrion 2 (2002) 87–93.

[16] R.P. Vazquez-Manrique, P. Gonzalez-Cabo, S. Ros, H. Aziz, H.A. Baylis, F. Palau, Reduction of *Caenorhabditis elegans* frataxin increases sensitivity to oxidative stress, reduces lifespan, and causes lethality in a mitochondrial complex II mutant, FASEB J. 20 (2006) 172–174.

[17] R.C. Lee, R.L. Feinbaum, V. Ambros, The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*, Cell 75 (1993) 843–854.

[18] S.Y. Ying, S.L. Lin, Intron-derived microRNAs—Fine tuning of gene functions, Gene 342 (2004) 25–28.

[19] T. Blumenthal, K. Steward, RNA processing and gene structure, in: D. Riddle, T. Blumenthal, B. Meyer, J. Priess (Eds.), *C. elegans* II, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1997, pp. 117–145.

[20] Z. Zhao, J.A. Sheps, V. Ling, L.L. Fang, D.L. Baillie, Expression analysis of ABC transporters reveals differential functions of tandemly duplicated genes in *Caenorhabditis elegans*, J. Mol. Biol. 344 (2004) 409–417.

[21] Y.H. Shim, J.J. Bonner, T. Blumenthal, Activity of a *C. elegans* GATA transcription factor, ELT-1, expressed in yeast, J. Mol. Biol. 253 (1995) 665–676.

[22] M.J. Gutch, A.J. Flint, J. Keller, N.K. Tonks, M.O. Hengartner, The *Caenorhabditis elegans* SH2 domain-containing protein tyrosine phosphatase PTP-2 participates in signal transduction during oogenesis and vulval development, Genes Dev. 12 (1998) 571–585.

[23] E.B. Maryon, B. Saari, P. Anderson, Muscle-specific functions of ryanodine receptor channels in *Caenorhabditis elegans*, J. Cell Sci. 111 (Pt. 19) (1998) 2885–2895.

[24] C.C. Mello, J.M. Kramer, D. Stinchcomb, V. Ambros, Efficient gene transfer in *C. elegans*: Extrachromosomal maintenance and integration of transforming sequences, EMBO J. 10 (1991) 3959–3970.

[25] D.T. Stinchcomb, J.E. Shaw, S.H. Carr, D. Hirsh, Extrachromosomal DNA transformation of *Caenorhabditis elegans*, Mol. Cell Biol. 5 (1985) 3484–3496.

[26] S.G. Clark, X. Lu, H.R. Horvitz, The *Caenorhabditis elegans* locus *lin-15*, a negative regulator of a tyrosine kinase signaling pathway, encodes two different proteins, Genetics 137 (1994) 987–997.

[27] L.S. Huang, P. Tzou, P.W. Sternberg, The *lin-15* locus encodes two negative regulators of *Caenorhabditis elegans* vulval development, Mol. Biol. Cell. 5 (1994) 395–411.

[28] M. Treinin, M. Chalfie, A mutated acetylcholine receptor subunit causes neuronal degeneration in *C. elegans*, Neuron 14 (1995) 871–877.

[29] J. Furst, M. Ritter, J. Rudzki, J. Danzl, M. Gschwentner, E. Scandella, M. Jakab, M. Konig, B. Oehl, F. Lang, P. Deetjen, M. Paulmichl, ICln ion channel splice variants in *Caenorhabditis elegans*: Voltage dependence and interaction with an operon partner protein, J. Biol. Chem. 277 (2002) 4435–4445.

[30] S. Eichmuller, V. Vezzoli, C. Bazzini, M. Ritter, J. Furst, M. Jakab, A. Ravasio, S. Chwatal, S. Dossena, G. Botta, G. Meyer, B. Maier, G. Valenti, F. Lang, M. Paulmichl, A new gene-finding tool: Using the *Caenorhabditis elegans* operons for identifying functional partner proteins in human cells, J. Biol. Chem. 279 (2004) 7136–7146.

[31] M.E. Hemler, Tetraspanin functions and associated microdomains, Nat. Rev. Mol. Cell Biol. 6 (2005) 801–811.

[32] M.G. Claros, P. Vincens, Computational method to predict mitochondrially imported proteins and their targeting sequences, Eur. J. Biochem. 241 (1996) 779–786.

[33] Y. Kiniwa, T. Fujita, M. Akada, K. Ito, T. Shofuda, Y. Suzuki, A. Yamamoto, T. Saida, Y. Kawakami, Tumor antigens isolated from a patient with vitiligo and T-cell-infiltrated melanoma, Cancer Res. 61 (2001) 7900–7907.

[34] W.K. Thomas, A.C. Wilson, Mode and tempo of molecular evolution in the nematode *Caenorhabditis*: Cytochrome oxidase II and calmodulin sequences, Genetics 128 (1991) 269–279.

[35] Y. Xue, H. Fares, B. Grant, Z. Li, A.M. Rose, S.G. Clark, E.Y. Skolnik, Genetic analysis of the myotubularin family of phosphatases in *Caenorhabditis elegans*, J. Biol. Chem. 278 (2003) 34380–34386.

[36] S. Brenner, The genetics of *Caenorhabditis elegans*, Genetics 77 (1974) 71–94.

[37] J.A. Lewis, J.T. Fleming, Basic Culture Methods, in: H.F. Epstein, D.C. Shakes (Eds.), *Caenorhabditis elegans*: Modern Biological Analysis of an Organism, Academic Press, San Diego, 1995, pp. 4–29.

[38] E.M. Southern, Detection of specific sequences among DNA fragments separated by gel electrophoresis, J. Mol. Biol. 98 (1975) 503–517.

[39] M.A. Frohman, Rapid amplification of cDNA ends, in: C.W. Dieffenbach, G.S. Dveksler (Eds.), PCR Primer: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1995, pp. 381–410.

[40] O. Hobert, PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic *C. elegans*, BioTechniques 32 (2002) 728–730.

[41] C. Mello, A. Fire, DNA transformation, Methods Cell Biol. 48 (1995) 451–482.

[42] J.M. Kramer, R.P. French, E.C. Park, J.J. Johnson, The *Caenorhabditis elegans rol-6* gene, which interacts with the *sqt-1* collagen gene to determine organismal morphology, encodes a collagen, Mol. Cell Biol. 10 (1990) 2081–2089.